PATENT ABSTRACTS OF JAPAN

(11)Publication number : **2001-117582**

(43)Date of publication of application : **27.04.2001**

(51)Int.Cl. **G10L 15/14**

G10K 15/04

G10L 13/00

G10L 15/00

G10L 15/10

// G10L101:04

G10L101:08

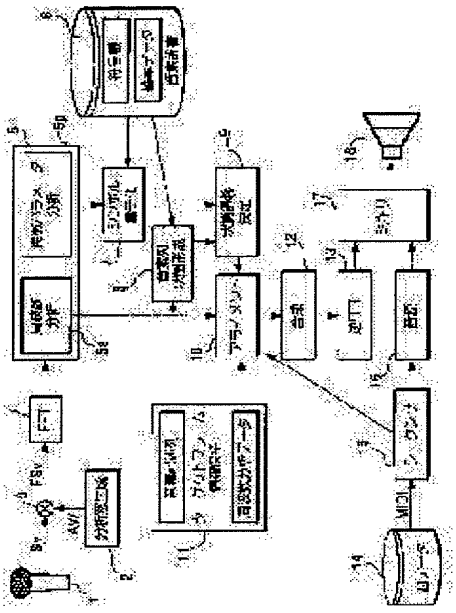
(21)Application number : **11-300276**

(22)Date of filing : **21.10.1999**

(71)Applicant : **YAMAHA CORP**
UNIV POMPEU FABRA

(72)Inventor : **KAWASHIMA TAKAHIRO**
PEDRO KEINO
ALEX ROSUKOSU

(54) VOICE PROCESSOR AND KARAOKE DEVICE



(57)Abstract:

PROBLEM TO BE SOLVED: To perform a voice process for making an object voice to be aligned correspond to an input voice in real time with a small storage capacity.

SOLUTION: Data obtained by analyzing an object voice in frame unit sectioned by specific time unit are stored and an input voice is also analyzed in frame unit sectioned by similar time unit. A hidden Markov model is generated according to a phoneme dictionary and the frame of an object person corresponding to the time of a frame of the input voice is specified in phoneme unit by using one-pass viterbi algorithm to match time relation. Data corresponding to the frame wherein the phonemes of the input voice and object voice are made to match with each other are composed.

Detailed Descriptions of the Invention:

.....

[0004]

The present invention is aimed at resolving said problem and the object thereof is to provide a voice processing apparatus capable of executing real-time processing with a small storage capacity for voice processing of associating in time series a target voice with an input voice for temporal alignment, and a karaoke apparatus having the voice processing apparatus.

.....

[0059]

As explained above, according to the present invention, it becomes possible to perform voice processing for associating in time series a target voice with an input voice, for the temporal alignment, using a small amount of storage capacity in the real-time processing.

(19)日本国特許庁（ＪＰ）　(12)**公開特許公報（Ａ）**　(11)特許出願公開番号
特開2001－117582
（P2001－117582A）
(43)公開日　平成13年４月27日(2001.4.27)

(51)Int.Cl.[7]	識別記号	ＦＩ		テーマコード*（参考）
Ｇ１０Ｌ　15/14		Ｇ１０Ｋ　15/04	３０２Ｄ	５Ｄ０１５
Ｇ１０Ｋ　15/04	３０２	Ｇ１０Ｌ 101:04		５Ｄ０４５
Ｇ１０Ｌ　13/00		101:08		５Ｄ１０８
15/00		3/00	５３５Ａ	
15/10			Ｑ	

審査請求　未請求　請求項の数20　ＯＬ　（全 18 頁）　最終頁に続く

(21)出願番号　特願平11－300276

(22)出願日　平成11年10月21日(1999.10.21)

(71)出願人　000004075
ヤマハ株式会社
静岡県浜松市中沢町10番１号

(71)出願人　599149108
ポンペウ　ファブラ大学
スペイン　バルセロナ　08002　メルセ12

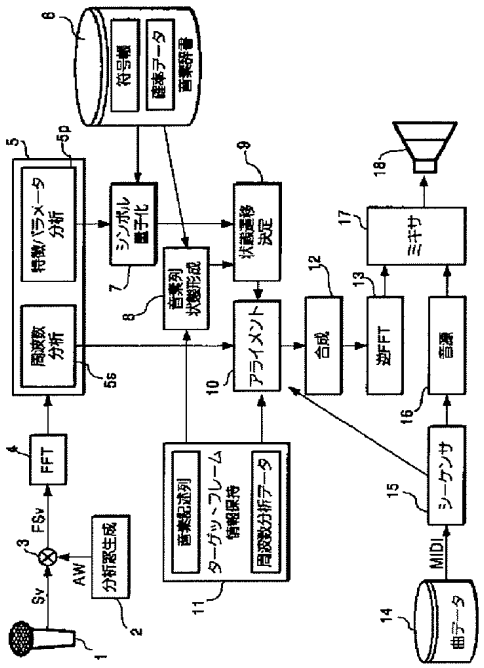
(72)発明者　川嶋　隆宏
静岡県浜松市中沢町10番１号　ヤマハ株式会社内

(74)代理人　100098084
弁理士　川▲崎▼　研二

最終頁に続く

(54)【発明の名称】　音声処理装置およびカラオケ装置

(57)【要約】
【課題】　アライメント対象となる対象音声と入力音声とを時系列で対応づける音声処理を、少ない記憶容量でリアルタイムで処理可能とする。
【解決手段】　所定の時間単位で区切ったフレーム単位で対象音声を分析したデータを記憶しておき、入力音声も同様の時間単位で区切ったフレーム単位で分析する。音素辞書に基づいて隠れマルコフモデルを形成し、１パスビタービアルゴリズムを用いて入力音声のフレームの時間に対応する対象者のフレームを音素単位で特定して時間関係を一致させる。そして、入力音声と対象音声の音素を一致させたフレームに対応するデータを合成する。



1

【特許請求の範囲】

【請求項１】　アライメント対象となる対象音声と入力音声とを時系列で対応づける音声処理装置であって、

前記対象音声をフレーム単位で分析し、前記対象音声の音素の時系列情報である音素列を複数の前記フレームから構成されるリージョン単位で記憶する対象音声情報記憶手段と、

音声信号の代表的な特徴パラメータを特徴ベクトルとして所定数のシンボルにクラスタ化した符号帳と、各音素毎に状態遷移確率および前記各シンボルの観測確率とを記憶する音素情報記憶手段と、

入力音声信号をフレーム単位で特徴パラメータ分析し、前記音素情報記憶手段に記憶された符号帳に基づいて前記入力音声の特徴パラメータをシンボル量子化して前記入力音声の観測シンボルとする入力音声量子化手段と、

前記音素情報記憶手段に記憶された状態遷移確率および観測確率に基づいて、前記音素列の各状態を有限状態ネットワーク上で隠れマルコフモデルによって形成する状態形成手段と、

前記入力音声量子化手段によって量子化された観測シンボルと、形成された前記隠れマルコフモデルに従って、ビタービアルゴリズムによって状態遷移を決定する状態遷移決定手段と、

決定した前記入力音声の状態遷移に基づいて前記入力音声の音素列と前記対象音声の音素列とを対応づけるアライメント手段とを備えることを特徴とする音声処理装置。

【請求項２】　請求項１記載の音声処理装置において、

前記特徴ベクトルは、音声のスペクトル特性をメルケプストラム係数で特徴づけるベクトルを含むことを特徴とする音声処理装置。

【請求項３】　請求項１記載の音声処理装置において、

前記特徴ベクトルは、音声のスペクトル特性を差分メルケプストラム係数で特徴づけるベクトルを含むことを特徴とする音声処理装置。

【請求項４】　請求項１記載の音声処理装置において、

前記特徴ベクトルは、音声を差分エネルギーで特徴づけるベクトルを含むことを特徴とする音声処理装置。

【請求項５】　請求項１記載の音声処理装置において、

前記特徴ベクトルは、音声をエネルギーで特徴づけるベクトルを含むことを特徴とする音声処理装置。

【請求項６】　請求項１記載の音声処理装置において、

前記特徴ベクトルは、音声の有声音らしさをゼロクロス率およびピッチエラーで特徴づけるベクトルを含むことを特徴とする音声処理装置。

【請求項７】　請求項１記載の音声処理装置において、

前記音素情報記憶手段は、大量の学習セットの予測ベクトルからクラスタ化アルゴリズムによってベクトル量子化することによって生成された符号帳を記憶することを特徴とする音声処理装置。

2

【請求項８】　請求項１記載の音声処理装置において、

前記音素情報記憶手段は、学習データに対するモデル尤度を最大にするパラメータを推定することによって求められた各音素における特徴ベクトルに対する状態遷移確率および観測シンボル確率を記憶することを特徴とする音声処理装置。

【請求項９】　請求項１記載の音声処理装置において、

前記状態遷移決定手段は、前記状態形成手段によって形成された状態から数状態前後の範囲から最適状態を検索して当該最適状態へ飛び越しを行うことを特徴とする音声処理装置。

【請求項１０】　請求項１記載の音声処理装置において、

前記状態形成手段は、前記対象音声の音素列にかかわらず、音素から無音状態あるいは息継ぎ状態への飛び越しを認めるパスを有する状態を形成することを特徴とする音声処理装置。

【請求項１１】　請求項１記載の音声処理装置において、

前記状態形成手段は、前記対象音声の音素列のうち近似する音素を有するグループについては、遷移確率を等価としたパスを有する状態を形成することを特徴とする音声処理装置。

【請求項１２】　請求項１記載の音声処理装置において、

前記アライメント手段は、前記入力音声の音素列に対応するフレームと前記対象音声の音素列に対応する前記リージョンを構成するフレームとを一致させることを特徴とする音声処理装置。

【請求項１３】　請求項１２記載の音声処理装置において、

前記アライメント手段は、決定した前記入力音声の音素に対応するフレーム数が、当該音素と一致する前記対象音声の音素に対応する前記リージョンを構成するフレーム数よりも多い場合には、予め記憶した所定フレームを用いて前記対象音声のフレーム数の不足フレーム数を補間することを特徴とする音声処理装置。

【請求項１４】　請求項１２記載の音声処理装置において、

前記アライメント手段は、決定した前記入力音声の音素に対応するフレーム数が、当該音素と一致する前記対象音声の音素に対応する前記リージョンを構成するフレーム数よりも少ない場合には、前記状態遷移決定手段によって前記入力音声の状態が遷移したと決定されたときに、次の音素の属する前記対象音声の前記リージョンを構成するフレームと一致させることを特徴とする音声処理装置。

【請求項１５】　請求項１記載の音声処理装置において、

前記状態遷移決定手段は、前記アライメント手段が摩擦

音をアライメントしている場合には、摩擦音あるいは前記対象音声の音素記述において次に位置する音素のみを考慮して状態遷移を決定することを特徴とする音声処理装置。

【請求項１６】　請求項１記載の音声処理装置において、

前記アライメント手段による対応に基づいて、前記対象音声と前記入力音声とを対応つけたフレーム単位で合成する合成手段を備えることを特徴とする音声処理装置。

【請求項１７】　請求項１６記載の音声処理装置において、

前記入力音声信号をフレーム単位で周波数分析して正弦波成分および残差成分を抽出する周波数分析手段を備え、

前記対象音声情報記憶手段は、前記対象音声を予めフレーム単位で周波数分析して抽出した正弦波成分および残差成分を記憶し、

前記合成手段は、前記入力音声の正弦波成分および残差成分と前記対象音声の正弦波成分および残差成分とを所定の割合で合成することを特徴とする音声処理装置。

【請求項１８】　請求項１７記載の音声処理装置において、

前記合成手段によって合成された正弦波成分および残差成分から逆周波数変換によって合成音声波形を生成する波形生成手段を備えることを特徴とする音声処理装置。

【請求項１９】　請求項１記載の音声処理装置を備えるカラオケ装置であって、

楽曲を構成する曲データを記憶する曲データ記憶手段と、

前記曲データに基づいて楽曲を再生する再生手段と、

前記対象音声の音素列および前記対象音声を予めフレーム単位で周波数分析して抽出したフレームデータを前記曲データの時系列と同期させる同期手段と、

前記アライメント手段による対応に基づいて、前記対象音声と前記入力音声とを対応つけたフレーム単位で合成する合成手段と、

前記同期手段による同期に基づいて、前記再生手段による楽曲の再生と前記合成手段によって合成された音声とを同期して出力する出力手段とを備えることを特徴とするカラオケ装置。

【請求項２０】　請求項１９記載のカラオケ装置において、

前記状態遷移決定手段は、前記曲データの時系列と同期して前記状態遷移確率への重み付け関数を付与することを特徴とするカラオケ装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】この発明は、アライメント対象となる対象音声と入力音声とを時系列で対応づける音声処理装置および当該音声処理装置を備えるカラオケ装置に関する。

【０００２】

【従来の技術】従来より、カラオケ等の技術分野において、歌唱者の歌声を歌手などの特定の歌唱者の歌声に似せて変換するといった音声処理技術が提案されている。通常、このような音声処理においては、二つの音声信号を時系列で対応づけるアライメントを行う必要がある。例えば図１に示すように、対象者が「なきながら」と発声した音声に歌唱者が「なきながら」と発声した音声を似せるように合成する場合であっても、対象者が「き」の音を発声するタイミングと歌唱者が「き」の音を発声するタイミングは異なる場合がある。このように人間が同じ言語を発音しても、その継続時間が異なる上に非線形に伸縮することが多いので、二つの音声を比較する場合には、同じ音素同士が対応するように、時間軸を非線形に伸縮する時間正規化するＤＰマッチング手法（Dynamic Time Warping:ＤＴＷ）という技術が知られている。ＤＰマッチング手法では、単語や音素に関して標準的な時系列を標準パターンとして用いているので、時系列パターンの時間的構造変化に対して音素単位で一致させることができる。またスペクトル変動に対して優れている隠れマルコフモデル（Hidden MarkovModel：ＨＭＭ）を用いる技術も知られている。隠れマルコフモデルでは、スペクトル時系列の統計的変動をモデルのパラメータに反映するので、話者の個人差などに起因するスペクトル変動に対して音素単位で一致させることができる。

【０００３】

【発明が解決しようとする課題】しかしながら、上述したＤＰマッチング手法を用いた場合には、スペクトル変動に対しては精度が悪く、従来の隠れマルコフモデルを用いる技術では記憶容量や計算量が膨大になるので、カラオケ装置における物まねなどのリアルタイム性が必要な処理には不向きだった。

【０００４】本発明は、上述した課題を解決するためになされたものであり、アライメント対象となる対象音声と入力音声とを時系列で対応づける音声処理を、少ない記憶容量でリアルタイム処理が可能な音声処理装置および当該音声処理装置を備えるカラオケ装置を提供することを目的としている。

【０００５】

【課題を解決するための手段】上述した課題を解決するために、請求項１に記載の発明は、アライメント対象となる対象音声と入力音声とを時系列で対応づける音声処理装置であって、前記対象音声をフレーム単位で分析し、前記対象音声の音素の時系列情報である音素列を複数の前記フレームから構成されるリージョン単位で記憶する対象音声情報記憶手段と、音声信号の代表的な特徴パラメータを特徴ベクトルとして所定数のシンボルにクラスタ化した符号帳と、各音素毎に状態遷移確率および

前記各シンボルの観測確率とを記憶する音素情報記憶手段と、入力音声信号をフレーム単位で特徴パラメータ分析し、前記音素情報記憶手段に記憶された符号帳に基づいて前記入力音声の特徴パラメータをシンボル量子化して前記入力音声の観測シンボルとする入力音声量子化手段と、前記音素情報記憶手段に記憶された状態遷移確率および観測確率に基づいて、前記音素列の各状態を有限状態ネットワーク上で隠れマルコフモデルによって形成する状態形成手段と、前記入力音声量子化手段によって量子化された観測シンボルと、形成された前記隠れマルコフモデルに従って、ビタービアルゴリズムによって状態遷移を決定する状態遷移決定手段と、決定した前記入力音声の状態遷移に基づいて前記入力音声の音素列と前記対象音声の音素列とを対応づけるアライメント手段とを備えることを特徴とする。

【０００６】請求項２に記載の発明は、請求項１記載の音声処理装置において、前記特徴ベクトルは、音声のスペクトル特性をメルケプストラム係数で特徴づけるベクトルを含むことを特徴とする。請求項３に記載の発明は、請求項１記載の音声処理装置において、前記特徴ベクトルは、音声のスペクトル特性を差分メルケプストラム係数で特徴づけるベクトルを含むことを特徴とする。請求項４に記載の発明は、請求項１記載の音声処理装置において、前記特徴ベクトルは、音声を差分エネルギーで特徴づけるベクトルを含むことを特徴とする。請求項５に記載の発明は、請求項１記載の音声処理装置において、前記特徴ベクトルは、音声をエネルギーで特徴づけるベクトルを含むことを特徴とする。請求項６に記載の発明は、請求項１記載の音声処理装置において、前記特徴ベクトルは、音声の有声音らしさをゼロクロス率およびピッチエラーで特徴づけるベクトルを含むことを特徴とする。

【０００７】請求項７に記載の発明は、請求項１記載の音声処理装置において、前記音素情報記憶手段は、大量の学習セットの予測ベクトルからクラスタ化アルゴリズムによってベクトル量子化することによって生成された符号帳を記憶することを特徴とする。請求項８に記載の発明は、請求項１記載の音声処理装置において、前記音素情報記憶手段は、学習データに対するモデル尤度を最大にするパラメータを推定することによって求められた各音素における特徴ベクトルに対する状態遷移確率および観測シンボル確率を記憶することを特徴とする。

【０００８】請求項９に記載の発明は、請求項１記載の音声処理装置において、前記状態遷移決定手段は、前記状態形成手段によって形成された状態から数状態前後の範囲から最適状態を検索して当該最適状態へ飛び越しを行うことを特徴とする。請求項１０に記載の発明は、請求項１記載の音声処理装置において、前記状態形成手段は、前記対象音声の音素列にかかわらず、音素から無音状態あるいは息継ぎ状態への飛び越しを認めるパスを有する状態を形成することを特徴とする。請求項１１に記載の発明は、請求項１記載の音声処理装置において、前記状態形成手段は、前記対象音声の音素列のうち近似する音素を有するグループについては、遷移確率を等価としたパスを有する状態を形成することを特徴とする。

【０００９】請求項１２に記載の発明は、請求項１記載の音声処理装置において、前記アライメント手段は、前記入力音声の音素列に対応するフレームと前記対象音声の音素列に対応する前記リージョンを構成するフレームとを一致させることを特徴とする。請求項１３に記載の発明は、請求項１２記載の音声処理装置において、前記アライメント手段は、決定した前記入力音声の音素に対応するフレーム数が、当該音素と一致する前記対象音声の音素に対応する前記リージョンを構成するフレーム数よりも多い場合には、予め記憶した所定フレームを用いて前記対象音声のフレーム数の不足フレーム数を補間することを特徴とする。請求項１４に記載の発明は、請求項１２記載の音声処理装置において、前記アライメント手段は、決定した前記入力音声の音素に対応するフレーム数が、当該音素と一致する前記対象音声の音素に対応する前記リージョンを構成するフレーム数よりも少ない場合には、前記状態遷移決定手段によって前記入力音声の状態が遷移したと決定されたときに、次の音素の属する前記対象音声の前記リージョンを構成するフレームと一致させることを特徴とする。請求項１５に記載の発明は、請求項１記載の音声処理装置において、前記状態遷移決定手段は、前記アライメント手段が摩擦音をアライメントしている場合には、摩擦音あるいは前記対象音声の音素記述において次に位置する音素のみを考慮して状態遷移を決定することを特徴とする。

【００１０】請求項１６に記載の発明は、請求項１記載の音声処理装置において、前記アライメント手段による対応に基づいて、前記対象音声と前記入力音声とを対応つけたフレーム単位で合成する合成手段を備えることを特徴とする。請求項１７に記載の発明は、請求項１６記載の音声処理装置において、前記入力音声信号をフレーム単位で周波数分析して正弦波成分および残差成分を抽出する周波数分析手段を備え、前記対象音声情報記憶手段は、前記対象音声を予めフレーム単位で周波数分析して抽出した正弦波成分および残差成分を記憶し、前記合成手段は、前記入力音声の正弦波成分および残差成分と前記対象音声の正弦波成分および残差成分とを所定の割合で合成することを特徴とする。請求項１８に記載の発明は、請求項１７記載の音声処理装置において、前記合成手段によって合成された正弦波成分および残差成分から逆周波数変換によって合成音声波形を生成する波形生成手段を備えることを特徴とする。

【００１１】請求項１９に記載の発明は、請求項１記載の音声処理装置を備えるカラオケ装置であって、楽曲を構成する曲データを記憶する曲データ記憶手段と、前記

曲データに基づいて楽曲を再生する再生手段と、前記対象音声の音素列および前記対象音声を予めフレーム単位で周波数分析して抽出したフレームデータを前記曲データの時系列と同期させる同期手段と、前記アライメント手段による対応に基づいて、前記対象音声と前記入力音声とを対応つけたフレーム単位で合成する合成手段と、前記同期手段による同期に基づいて、前記再生手段による楽曲の再生と前記合成手段によって合成された音声とを同期して出力する出力手段とを備えることを特徴とする。請求項２０に記載の発明は、請求項１９記載のカラオケ装置において、前記状態遷移決定手段は、前記曲データの時系列と同期して前記状態遷移確率への重み付け関数を付与することを特徴とする。

【００１２】

【発明の実施の形態】以下、図面を参照しながら、本発明の実施の形態について説明する。

【０１３】[１．実施形態の構成]

[１－２．全体構成]図２は、本実施形態の構成を示す図である。実施形態は、本発明を物まね機能付きのカラオケ装置に適用したものであり、歌唱者（Me）のマイクからの入力音声を、例えば歌手などの物まね対象者（ターゲット：Target）の音声に似せるように音声変換を行って出力するように構成されている。より具体的には、所定の時間単位で区切ったフレーム単位で対象音声を分析したデータを記憶しておき、入力音声も同様の時間単位で区切ったフレーム単位で分析することにより、入力音声のフレームの時間に対応する対象者のフレームを特定できれば、時間関係を一致させることができるようになっている。そして、音素単位で入力音声と対象音声とを一致させたフレームデータを合成することによって音声変換を行うように構成されている。

【００１４】図２において、マイク１は、ものまねをしようとする歌唱者の声を収集し、入力音声信号Ｓｖとして入力音声信号切出部３に出力する。分析窓生成部２は、前回のフレームで検出したピッチの周期の固定倍の周期を有する分析窓（例えば、ハミング窓）ＡＷを生成し、入力音声信号切出部３に出力する。なお、初期状態あるいは前回のフレームが無声音（含む無音）の場合には、予め設定した固定周期の分析窓を分析窓ＡＷとして入力音声信号切出部３に出力する。入力音声信号切出部３は、入力された分析窓ＡＷと入力音声信号Ｓｖとを掛け合わせ、入力音声信号Ｓｖをフレーム単位で切り出し、フレーム音声信号ＦＳｖとして高速フーリエ変換部４に出力する。高速フーリエ変換部４は、フレーム音声信号ＦＳｖから周波数スペクトルを求め、周波数分析部５ｓおよび特徴パラメータ分析部５ｐを備えた入力音声分析部５に出力する。

【００１５】周波数分析部５ｓは、後述するＳＭＳ（Spectral Modeling Synthesis）分析を行って正弦波成分および残差成分を抽出し、分析した当該フレームの歌唱者の周波数成分情報として保持する。特徴パラメータ分析部５ｐは、入力音声のスペクトル特性を特徴づける特徴パラメータを抽出し、シンボル量子化部７に出力する。本実施形態では、特徴パラメータとして後に説明する５種類（メルケプストラム係数、差分メルケプストラム係数、差分エネルギー係数、エネルギー、ボイスネス）の特徴ベクトルを用いる。

【００１６】音素辞書記憶部６は、後に詳しく説明するように、符号帳および各音素における特徴ベクトルの状態遷移確率とシンボル発生確率とを示す確率データを含む音素辞書を記憶している。シンボル量子化部７は、音素辞書記憶部６に記憶された符号帳を参照して、そのフレームにおける特徴シンボルを選び出し、状態遷移決定部９に出力する。音素列状態形成部８は、隠れマルコフモデル（ＨＭＭ）によって音素列状態を形成し、状態遷移決定部９は、入力音声から得られたフレーム単位の特徴シンボルを用いて、後述するビタービアルゴリズムに従って状態遷移を決定する。

【００１７】アライメント部１０は、決定された状態遷移から入力音声の時間ポインタを決定し、当該時間ポインタに対応するターゲットフレームを特定し、周波数分析部に保持された入力音声の周波数成分と、ターゲットフレーム情報保持部１１に保持された対象者の周波数成分とを合成部１２に出力する。ターゲットフレーム情報保持部１１には、予めフレーム単位で周波数分析された周波数分析データおよび、いくつかのフレームで構成される時間リージョン（region）単位で記述された音素列が記憶されている。

【００１８】合成部１２は、入力音声の周波数成分と対象者の周波数成分とを所定の割合で合成した新たな周波数成分を生成して逆高速フーリエ変換部１３に出力し、逆高速フーリエ変換部１３は新たな周波数成分を逆高速フーリエ変換して新たな音声信号を生成する。ところで、本実施形態は物まね機能を備えたカラオケ装置であり、曲データ記憶部１４には、ＭＩＤＩデータや時間データ、歌詞データなどによって示されるカラオケ曲データが記憶されており、ＭＩＤＩデータを時間データに従って再生するシーケンサ１５およびシーケンサ１５の出力データから楽音信号を生成する音源１６を備えている。ミキサ１７は、逆高速フーリエ変換部１３から出力された音声信号と音源１６から出力された楽音信号とを合成してスピーカ１８から出力する。このように、歌唱者がマイク１に向かって歌唱すると、歌唱者の音声が対象者の音声に似せて変換された新たな音声と、カラオケの伴奏楽音とがスピーカ１８から出力されるように構成されている。

【００１９】[１－２．音素辞書]次に、本実施形態で用いる音素辞書について説明する。音素辞書は、音声信号の代表的な特徴パラメータを特徴ベクトルとして所定数のシンボルにクラスタ化した符号帳と、各音素毎に状態

遷移確率および前記各シンボルの観測確率とから構成されいる。

【００２０】[１－２－１．特徴ベクトル]符号帳について説明する前に、まず、本実施形態で用いる特徴ベクトルについて説明しておく。

①メルケプストラム係数（b_{MEL}）

メルケプストラム係数は、音のスペクトル特性を少ない次数で表す係数であり、本実施形態では１２次元ベクトルとして１２８シンボルにクラスタ化している。

②差分メルケプストラム係数（$b_{deltaMEL}$）

差分メルケプストラム係数は、メルケプストラム係数の時間差分を表す係数であり、本実施形態では１２次元ベクトルとして１２８シンボルにクラスタ化している。

③差分エネルギー係数（$b_{deltaENERGY}$）

差分エネルギー係数は、音の強さの時間差分を表す係数であり、本実施形態では１次元ベクトルとして３２シン＊ボルにクラスタ化している。

④エネルギー（b_{ENERGY}）

エネルギーは、音の強さを表す係数であり、本実施形態では１次元ベクトルとして３２シンボルにクラスタ化している。

⑤ボイスネス（$b_{VOICENESS}$）

ボイスネスは、有声音らしさを表す特徴ベクトルであり、音声をゼロクロス率およびピッチエラーで特徴づける２次元ベクトルとして３２シンボルにクラスタ化している。以下、ゼロクロス率とピッチエラーについてそれぞれ説明する。

【００２１】（１）ゼロクロス率

ゼロクロス率は、有声音であるほどゼロクロス率が低くなる特徴を有するものであり、次式で定義される。

【数１】

$$Z_s(n)=\frac{1}{N}\sum_{m=n-M+1}^{n}\frac{\left|\mathrm{sgn}\{x(m)\}-\mathrm{sgn}\{x(m-1)\}\right|}{2}w(n-m)$$

ここで、sgn{s(n)}-+1:s(n)>-0,-1:s(n)<0,

N：フレームサンプル数

W：フレーム窓

s：入力信号

【００２２】（２）ピッチエラー

ピッチエラーは、予測ピッチから測定ピッチへのエラーおよび、測定ピッチから予測ピッチへのエラーの２方向からのミスマッチを求めることによって有声音らしさを示すものであり、詳細には、"Fundamental Frequency Estimation in the SMS Analysis"(P.Cano. Proceedings of the Digital Audio Effects Workshop,1898) にTwo Way Mismuch手法として説明されている。

【００２３】まず、予測ピッチ(p)から測定ピッチ(m)へのピッチエラーは次式で表される。

【数２】

$$Err_{p\to m}=\sum_{n=1}^{N}E_w(\Delta f_n,f_n,a_n,A_{\max})$$
$$=\sum_{n=1}^{N}\left\{\Delta f_n\cdot(f_n)^{-p}+\left(\frac{a_n}{A_{\max}}\right)\times\left[q\Delta f_n\cdot(f_n)^{-p}-r\right]\right\}$$

fn：ｎ番目の予測ピーク周波数

Δfn：ｎ番目の予測ピーク周波数とそれに近接した測定ピーク周波数差

a_n：ｎ番目の測定アンプリチュード

Amax：最大アンプリチュード

【００２４】一方、測定ピッチ(m)から予測ピッチ(p)へのピッチエラーは次の式で表される。

【数３】

$$Err_{m\to p}=\sum_{k=1}^{N}E_w(\Delta f_k,f_k,a_k,A_{\max})$$
$$=\sum_{k=1}^{N}\left\{\Delta f_k\cdot(f_k)^{-p}+\left(\frac{a_k}{A_{\max}}\right)\times\left[q\Delta f_k\cdot(f_k)^{-p}-r\right]\right\}$$

fk：ｋ番目の予測ピーク周波数

Δfk：ｋ番目の予測ピーク周波数とそれに近接した測定ピーク周波数差

ak：ｋ番目の測定アンプリチュード

Amax：最大アンプリチュード

【００２５】従って、トータルエラーは次式のようになる。

【数４】

$$Err_{total}=Err_{p\to m}/N+\rho Err_{m\to p}/K$$

なお、常数として、p=0.5,q=1.4,r=0.5が実験的にほとんどの音声に対して最適であることが報告されている。

【００２６】[１－２－２．符号帳]符号帳は、それぞれの特徴ベクトルに対して、各シンボルの数へクラスタされたベクトル情報が記憶されている（図３参照）。符号帳は、大量の学習セット中の全ての予測ベクトルの中から、最小歪みである量子化によって、Ｋ予測ベクトル（コード）と言われるセットを見つけることによって作成されている。本実施形態では、クラスタ化のアルゴリズムとしてＬＧＢアルゴリズムを用いる。

【００２７】以下、ＬＧＢアルゴリズムを以下に示す。

①初期化

まず、ベクトルの全体の中からセントロイドを見つける。ここでは、初期コードベクトルとする。

②反復

11

Ｉをトータル反復回数とすると、2^Iのコードベクトルが要求される。そこで、反復回数をｉ＝１，２，……，Ｉとすると、反復ｉについて、以下の計算を行う。

１）いくつかの存在するｘというコードベクトルを、ｘ（１＋ｅ）とｘ（１－ｅ）という二つのコードへ分割する。ここで、ｅは、例えば０．００１という小さな数値である。これにより、２ｉ個の新しいコードベクトルx^i_k（ｋ＝１，２，…，2^i）が得られる。

２）学習セット中の各々の予測ベクトルｘについて、ｘからコードへX^i_k量子化する。

$k' = argmin_k d(x, x^i_k)$

ここで、ｄ（ｘ、x^i_k）は、予測空間での歪み距離を示している。

３）反復計算の間、各々のｋについて、$x^i_k = Q(x)$のように、すべてのベクトルをセントロイドする計算を行う。

【００２８】[１－２－３．確率データ]次に、確率データについて説明する。本実施形態では、音声をモデル化するためのサブワード単位としてＰＬＵ（疑似音素単位）を用いる。より具体的には、図４に示すように、日本語を２７の音素単位で扱うものとし、各音素には状態数が割り付けられている。状態数とは、サブワード単位の持続する最も短いフレーム数をいい、例えば、音素“ａ”の状態数は“３”であるので、音素“ａ”は少なくとも３フレーム続くことを意味する。３状態は、発音の立ち上がり・定常状態・リリース状態を擬似的に表したものである。音素“ｂ”や“ｇ”などの破裂音は、本来持つ音韻が短いので２状態に設定されており、息継ぎ（ASPIRATION）も２状態に設定されている。そして、無音（SILEMCE）は、時間的変動がないので１状態に設定されている。

【００２９】図５に示すように、音素辞書中の確率データには、サブワード単位で表される２７の音素に対して、各状態の遷移確率と、各特徴ベクトルのシンボルに対する観測シンボル発生確率が記述されている。なお、図５においては、記載を中略しているが、各特徴ベクトル毎の観測シンボル発生確率の和は１となっている。これらのパラメータは、学習データに対するモデルの尤度を最大にするサブワード単位モデルのパラメータを推定することにより求める。ここでは、セグメントｋ平均学習アルゴリズムを用いる。

【００３０】セグメントｋ平均学習アルゴリズムを以下に示す。

①初期化

まず、予め音素セグメント分けされた初期推定データについて、各々の音素セグメントをＨＭＭ状態へ線形的にセグメント化（分割）する。

②推定

遷移確率は、次式に示すように、遷移に用いられる遷移数（フレーム単位）をカウントし、これを、状態からの

12

遷移全てに用いられる遷移数（フレーム単位）のカウント値で割り算することにより求められる。

【数５】

$$\hat{a}_{ij} = \frac{Siから Sjまでの遷移数}{Siからの遷移数}$$

【００３１】一方、シンボル発声確率は、次式に示すように、各状態で各特徴シンボルを発生する数をカウントし、これを各状態における全ての発生数のカウントで割り算することによって求められる。

【数６】

$$\hat{b}_j(O_k) = \frac{Siでの特徴シンボル O_k の時間数}{S_j での時間数}$$

【００３２】③セグメンテーション

学習セットに対して、ステップ②で求めた推定パラメータを用いて、ビタービアルゴリズムを介して再セグメント化する。

④反復

ステップ②とステップ③を収束するまで繰り返す。

【００３３】[１－３．ターゲットフレーム情報]ターゲットフレーム情報保持部１１には、予め対象者の音声がＳＭＳ分析されてフレーム単位で記憶されている。まず、図６を参照しながら、ＳＭＳ分析について説明する。ＳＭＳ分析では、まず標本化された音声波形に窓関数を乗じた音声波形（Frame）を切り出し、高速フーリエ変換（FFT）を行って得られる周波数スペクトルから、正弦波成分と残差成分とを抽出する。

【００３４】正弦波成分とは、基本周波数（Pitch）および基本周波数の倍数にあたる周波数（倍音）の成分をいう。本実施形態では、基本周波数を“Fi”として保持し、各成分の平均アンプリチュードを“Ai”として保持し、スペクトル包絡をエンベロープとして保持する。残差成分とは、入力信号から正弦波成分を除いた成分であり、本実施形態では、図６に示すように周波数領域のデータとして保持する。図６に示すように得られた正弦波成分および残差成分で示される周波数分析データは、図７に示すようにフレーム単位で記憶される。本実施形態では、フレーム間の時間間隔は５ｍｓとし、フレームをカウントすることによって時間を特定することができるようになっている。各フレームには曲の冒頭からの経過時間に相当するタイムスタンプが付されている（ｔｔ１、ｔｔ２、……）。

【００３５】ところで、先に説明したように、各音素は、少なくとも音素毎に設定されている状態数分のフレームが続くから、ターゲットフレーム情報においても、各音素情報は複数のフレームから構成される。この複数フレームのまとまりをリージョン（region）とする。ターゲットフレーム情報保持部には、対象者が歌唱したときの音素列が記憶されるが、各音素とリージョンとを対

応つけて記述している。図７に示す例では、フレームｔｔ１～ｔｔ５から構成されるリージョンが音素“ｎ”に対応し、フレームｔｔ６～ｔｔ１０から構成されるリージョンが音素“ａ”に対応している。このように、ターゲットフレーム情報を保持し、同様のフレーム分析を入力音声に対して行えば、音素単位で両者を一致させた際に、フレームで時間を特定することができ、周波数分析データで合成処理ができるようになる。

【００３６】[２．実施形態の動作]次に、実施形態の動作について説明する。

【００３７】[２－１．概要動作]最初に、概要動作について図８に示すフローチャートを参照しながら説明する。まず、マイク入力音声分析が行われる（Ｓ１）。具体的には、フレーム単位で高速フーリエ変換し、周波数スペクトルからＳＭＳ分析を行った周波数分析データを保持する。また、周波数スペクトルから特徴パラメータ解析を行って、音素辞書に基づいてシンボル量子化を行う。

【００３８】次に、音素辞書および音素記述列に基づいて、HMMモデルによる音素の状態決定を行い（Ｓ２）、シンボル量子化された特徴パラメータおよび決定された音素状態に基づいて１パスビタービアルゴリズムによって状態遷移を決定する（Ｓ３）。HMMモデルおよび１パスビタービアルゴリズムについては後に詳しく説明する。そして、決定した状態遷移により入力音声の時間ポインタを決定し（Ｓ４）、当該時間が新たな音素状態になったか否かを判定する（Ｓ５）。時間ポインタとは、入力音声および対象音声の時系列において、当該処理時刻におけるフレームを特定するものである。本実施形態では、入力音声および対象音声はフレーム単位で周波数分析され、各フレームは、入力音声および対象音声の時系列と対応付けられている。以後、入力音声の時系列を時刻ｔｍ１、ｔｍ２……と表記し、対象音声の時系列をｔｔ１、ｔｔ２……と表記する。

【００３９】ステップＳ５の判定において、新たな音素状態になったと判定した場合は（Ｓ５；Ｙｅｓ）、フレームカウントを開始し（Ｓ６）、時間ポインタを音素列の先頭へ移動する（Ｓ７）。フレームカウントとは、当該音素状態として処理したフレーム数をいい、先に説明したように、各音素は複数のフレームが続くので、すでに何フレーム続いたかを示す値となる。そして、入力音声フレームの周波数分析データと対象者音声フレームの周波数分析データとを周波数領域で合成し（Ｓ８）、逆高速フーリエ変換することによって（Ｓ９）新たな音声信号を生成して出力する。

【００４０】ところで、ステップＳ５の判定において、新たな音素状態に遷移していないと判定した場合は（Ｓ５；Ｎｏ）、フレームカウントをインクリメントして（Ｓ１０）、時間ポインタをフレーム時間間隔分進め（Ｓ１１）、ステップＳ８に移行する。具体例をあげて

説明すると、図７に示す例では、音素状態が“ｎ”にとどまり続ける場合はフレームカウントをインクリメントして、時間ポインタをｔｔ１、ｔｔ２……と移動させる。しかし、フレームｔｔ３の音素状態が“ｎ”を処理した次の時刻に“ａ”に遷移した場合には、音素列“ａ”の先頭フレームｔｔ６に時間ポインタを移動させる。このようにすれば、対象者と歌唱者との発音タイミングが異なっても、音素単位での時間一致を行うことができる。

【００４１】[２－２．動作の詳細]次に、概要動作においてふれた各処理について詳細に説明する。

【００４２】[２－２－１．入力音声分析]図９は、入力音声を分析する処理について詳細に説明する図である。図９に示すように、入力音声波形からフレーム単位で切り出された音声信号は、高速フーリエ変換によって周波数スペクトルに変換される。周波数スペクトルは、先に説明したＳＭＳ分析によって周波数成分データとして保持される他、特徴パラメータ解析が行われる。一方、周波数スペクトルは、特徴パラメータ分析も行われる。より具体的には、各特徴ベクトル毎に、音素辞書から最大尤度のシンボルを見つけることによってシンボル量子化して観測シンボルとする。このようにして得られたフレーム毎の観測シンボルを用いて、後に詳しく説明するように状態遷移が決定されるようになる。

【００４３】[２－２－２．隠れマルコフモデル]次に、図１０を参照しながら、隠れマルコフモデル（HMM）について説明する。なお、音声の状態は一方向へ遷移するので、left to right型のモデルを用いている。

【００４４】時刻ｔにおいて、状態がｉからｊへ遷移する確率（状態遷移確率）をa_{ij}と表す。図１０に示す例では、状態①にとどまる確率をa_{11}と表し、状態①から状態②へ遷移する確率をa_{12}と表している。各状態の中には特徴ベクトルがそれぞれ存在し、各々に異なる観測シンボルがある。これを$X=\{x_1, x_2, \cdots, x_T\}$と表す。そして、時刻ｔにおいて状態がｊである時に特徴ベクトルのシンボルx_tを発生させる確率（観測シンボル離散確率）を$b_j(x_t)$と表す。モデルλにおいて、時刻Ｔまでの状態系列を$Q=\{q_1, q_2, \cdots, q_T\}$とすると、観測シンボル系列Ｘと状態系列Ｑの同時発生確率は、次式で表せる。

【数７】

$$P(X,Q\backslash\lambda)=a_{q1q2}\prod_{t=1}^{T}b_{qt}(x_t)a_{qtq(t+1)}$$

観測シンボル系列は判っているが、状態系列は観測しえないという理由で、このようなモデルが隠れマルコフモデル（HMM）と呼ばれている。本実施形態では、ターゲットフレーム情報保持部１１に記憶されている音素記述列に基づいて、図１０に示すようなＦＮＳ（有限状態ネットワーク）を音素単位で形成する。

15

【００４５】[２－２－３．アライメント]次に、図１１および図１２を参照しながら、本実施形態におけるアライメントについて説明する。本実施形態では、音素記述列に基づいて形成された上述の隠れマルコフモデルと、入力音声から抽出したフレーム単位の特徴シンボルを用いて、１パスビタービアルゴリズムに従って入力音声の状態遷移を決定する。そして、入力音声の音素と対象音声の音素とをフレーム単位で対応づける処理を行う。また、本実施形態では、二つの音声信号のアライメントをカラオケ装置において用いているので、曲データに従った楽曲の時系列と、音声信号の時系列とを同期させる処理も行う。以下、これらの処理について順次説明する。

【００４６】[２－２－３－１．１パスビタービアルゴリズム]ビタービアルゴリズムは、観測シンボル系列の

16

＊各観測シンボルが各ＨＭＭモデルによって出現する全ての確率を算出し、最大確率を与えるパスを後から選択して状態遷移結果とするものである。しかしながら、観測シンボル系列が終結した後に状態遷移結果を求めるので、リアルタイム処理には不向きである。そこで、本実施形態では、以下に説明する１パスビタービアルゴリズムを用いて、その時点まで音素状態を決定する。下記式におけるΨ_t（ｊ）は、時刻ｔフレームまでの観測をふまえて算出した、一つのパスを経由して得られる時刻ｔのフレームにおけるベスト確率δ_t（ｉ）を最大とする状態を選択する。すなわち、Ψ_t（ｊ）に従って音素状態が遷移していく。初期演算としてδ_1（ｉ）＝１とし、繰り返し演算として

【数８】

$$\delta_t(j)=\max_{j-1<i<j}\left[\delta_{t-1}(i)a_{ij}\right]\cdot b_{j(MEL)}(O_t)\cdot b_{j(deltaMEL)}(O_t)\cdot b_{j(deltaENERGY)}(O_t)\cdot b_{j(VOICENESS)}(O_t)\cdot b_{j(ENERGY)}(O_t)$$

$$1\le t\le T,\ 1\le j\le t$$

$$\psi_t(j)=\arg\max_{j-1<i<j}\left[\delta_{t-1}(i)a_{ij}\right]$$

$$1\le t\le T,\ 1\le j\le t$$

を実行する。ここで、a_{ij}は状態ｉから状態ｊへの状態遷移確率であり、b_j（O_t）は特徴ベクトルの時刻ｔにおけるシンボル発生確率である。各観測シンボルは、入力音声から抽出された特徴ベクトルであるから、歌唱者の発声態様によって観測シンボルが異なり、遷移の態様も異なるようになる。図１１に示す例では、上記式によって計算された確率を○あるいは△で示している（○＞△）。例えば、時刻ｔｍ１から時刻ｔｍ３までの観測をふまえ、状態“Ｓｉｌｅｎｃｅ”から状態“ｎ１”へのパスが形成される確率は、状態“Ｓｉｌｅｎｃｅ”から状態“Ｓｉｌｅｎｃｅ”へのパスが形成される確率よりも高く、時刻ｔｍ３におけるベスト確率となり、図中太矢印で示すように状態遷移を決定する。このような演算を入力音声の各フレームに対応する時刻（ｔｍ１、ｔｍ２、……）毎に行うことによって、図１１に示す例では、時刻ｔｍ３において状態“Ｓｉｌｅｎｃｅ”から状態“ｎ１”に遷移し、時刻ｔｍ５において状態“ｎ１”から状態“ｎ２”に遷移し、時刻ｔｍ９において状態“ｎ２”から状態“ｎ３”に遷移し、時刻ｔｍ１１において状態“ｎ３”から状態“ａ１”に遷移したように決定されている。これにより、入力音声の音素をフレーム単位の各時刻において特定できるようになる。

【００４７】[２－２－３－２．フレーム単位の対応]上述したように状態遷移を決定し、入力音声の音素がフレーム単位で特定されると、次に、特定された音素に対応する対象音声のフレームを特定する。上述したように、隠れマルコフモデルの各状態はターゲットフレーム情報保持部１１に記憶された対象音声の音素列記述に基づいて形成されているので、各状態に対応する対象音声の音素毎のフレームを特定することができるようになっている。本実施形態では、アライメントとして、対象音声と入力音声の対応する音素が同じフレーム同士を、各フレーム毎に時系列で一致させる処理を行う。

【００４８】図１１に示す例では、対象音声の時刻ｔｔ１～ｔｔ３のフレームが音素“Ｓｉｌｅｎｃｅ”に対応し、時刻ｔｔ４～ｔｔ９のフレームが音素“ｎ”に対応し、時刻ｔｔ１０～のフレームが音素“ａ”に対応している。一方、１パスビタービアルゴリズムによって入力音声の状態遷移が決定され、入力音声の時刻ｔｍ１～ｔｍ２のフレームが音素“Ｓｉｌｅｎｃｅ”に対応し、時刻ｔｍ３～ｔｍ１０のフレームが音素“ｎ”に対応し、時刻ｔｍ１１～のフレームが音素“ａ”に対応している。そして、音素“Ｓｉｌｅｎｃｅ”に対応するフレームとして、入力音声の時刻ｔｍ１のフレームと対象音声の時刻ｔｔ１のフレームを一致させ、入力音声の時刻ｔｍ２のフレームと対象音声の時刻ｔｔ２のフレームを一致させる。入力音声の時刻ｔｍ３において状態“Ｓｉｌｅｎｃｅ”から状態“ｎ１”に遷移しているので、音素“ｎ”に対応するフレームとしては、入力音声の時刻ｔｍ３のフレームが最初のフレームになる。一方、対象音声のフレームは、音素“ｎ”に対応するフレームは、音素列記述によれば時刻ｔｔ４のフレームからであるので、音素“ｎ”発声開始時の対象音声の時間ポインタは時刻ｔｔ４となる（図８：ステップＳ５～Ｓ７参照）。

次に、入力音声の時刻ｔｍ４においては、新たな音素状態に遷移していないので、フレームカウントをインクリメントするとともに、対象音声の時間ポインタをフレーム時間間隔分進めて（図８：ステップＳ５～Ｓ１１参照）、時刻ｔｔ５のフレームを入力音声の時刻ｔｍ４のフレームと一致させる。このようにして、入力音声の時刻ｔｍ５～ｔｍ７と、対象音声の時刻ｔｔ６～ｔｔ８とを順次一致させていく。

【００４９】ところで、図１１に示す例では、入力音声の時刻ｔｍ３～ｔｍ１０までの８フレーム分が音素“ｎ”に対応しているのに対して、対象音声の音素“ｎ”に対応しているフレームは時刻ｔｔ４～ｔｔ９までのフレームである。このように、歌唱者が対象者よりも同じ音素を長い時間発声してしまう場合が生じるので、本実施形態では、予め用意したループフレームを用いて対象音声が入力音声よりも短い場合の補間を行う。ループフレームは、音をのばして発音する場合のピッチの変化やアンプリチュードの変化を擬似的に再現するためのデータを数フレーム分記憶しており、例えば、基本周波数の差分（△Pitchi）やアンプリチュードの差分（△Amp）などから構成される。そして、ターゲットフレームデータ中には、音素列における各音素の最終フレームにループフレームの呼び出しを指示するデータを記述しておく。これにより、歌唱者が対象者よりも同じ音素を長い時間発声してしまった場合でも、良好にアライメントを行うことができるようになる。

【００５０】[２－２－３－３．曲データとの同期]ところで、本実施形態では、カラオケ装置に音声変換を適用しており、カラオケ装置はＭＩＤＩデータに基づいて楽曲の演奏を行うので、音声の進行と楽曲の進行が同期していることが望ましい。そこで、本実施形態では、アライメント部１０は、曲データで示される時系列と対象音声の音素列とを同期させるように構成している。より具体的には、図１２に例示するように、シーケンサ１５は曲データに記述された時間情報（例えば、ＭＩＤＩデータの再生時間間隔を示す△タイムやテンポ情報）などに基づいて、楽曲の進行情報を生成してアライメント部１０に出力する。アライメント部１０は、シーケンサ１５から出力された時間情報とターゲットフレーム情報保持部１１に記憶されている音素記述列とを比較して、曲進行の時系列と対象音声の時系列とを対応づける。

【００５１】また、図１２に示すような重み付け関数ｆ（｜$t_m - t_e$｜）を用いて、楽曲に同期して状態遷移確率への重み付けをおこなうことができるようにしている。この重み付け関数は、各状態遷移確率a_{ij}に乗じる窓関数である。なお、図中ａおよびｂは楽曲のテンポに応じた要素である。また、αは限りなく０に近い値に設定する。上述したように、対象音声の時間ポインタは楽曲のテンポに同期して進行するので、このような重みつけ関数を導入することによって、結果的に歌唱音声と対象音声との同期が正確になる。

【００５２】[３．変形例]本発明は、上述した実施形態に限定されるものではなく、以下に説明するような各種の変形が可能である。

【００５３】[３－１．音素の飛び越し]上記実施形態では、１パスビタービアルゴリズムを用いて状態遷移を決定しているが、歌唱者が歌詞を間違えた場合には不向きである。例えば、数フレーズ先の歌詞を歌ってしまった場合や、数フレーム前の歌詞を歌ってしまった場合などが考えられる。このような場合は、図１３に示すように、数状態前後まで最適状態を検索する範囲を広げ、最適状態と判断した場合に限り飛び越しを行うようにすればよい。より具体的には、入力音声の時刻ｔｍ４においては、音素“ａ”に対応する状態となっているので、上述した１パスビタービアルゴリズムによれば、入力音声の時刻ｔｍ５のフレームについては、音素“ａ”から遷移しない確率、あるいは音素列記述において音素“ａ”の次にくる“Ｓｉｌｅｎｃｅ”への遷移確率のいずれか高い方から最大確率を選択することになる。しかしながら、歌唱者は無音期間なしに音素“ｋ”の発声を開始しているので、対象者の音素列記述のうちの“Ｓｉｌｅｎｃｅ”については飛び越してアライメントすることが望ましい。そこで、このような歌唱者が対象者の音素列記述に従わずに発声した場合には、数状態前後の状態まで最大確率となる状態を検索するようにしてもよい。図１３に示す例では、直前のフレーム状態の前後３状態の範囲を検索して、２状態先の音素“ｋ”を最大確率としている。このようにて、“Ｓｉｌｅｎｃｅ”を飛び越して音素“ｋ”への状態遷移を決定する。

【００５４】また、無音の位置や息継ぎの位置などが異なる場合も多い。このような場合には、上記実施形態では音素の位置が異なってしまう。そこで、図１３に示すように、発音音素単位から“Ｓｉｌｅｎｃｅ”と“Ａｓｐｉｌａｔｉｏｎ”や発音音素単位への飛び越しの確率を同じように設定する。例えば、対象者の音素列記述においては、音素“ｉ”の前後数状態には“Ａｓｐｉｌａｔｉｏｎ”は記述されていない。しかしながら、音素記述列において音素“ｉ”の次に記述されている音素“ｎ”へ遷移する確率と、記述されていない“Ｓｉｌｅｎｃｅ”あるいは“Ａｓｐｉｌａｔｉｏｎ”への飛び越しを行う確率を同等に設定し、“Ｓｉｌｅｎｃｅ”あるいは“Ａｓｐｉｌａｔｉｏｎ”に飛び越しを行った後に、音素記述列中の音素に戻る確率も同等に設定しておけばよい。　このようにすれば、例えば図１３に示す例のように、歌唱者が時刻ｔｍ７において、対象者の音素記述列に従わずに息継ぎを行った場合でも柔軟にアライメントすることができる。また、対象者の音素列記述にかかわらず、ある摩擦音の次に他の摩擦音に遷移する場合があるので、摩擦音をアライメントしている時は、摩擦音あるいは対象音声の音素記述の次の音素について最

大確率を検索するようにしてもよい。

【００５５】[３－２．似通った音素]日本語では、同じ言葉でも歌唱者によって異なる音素で発音する場合がある。たとえば、図１４に示すように、音素記述では“nagara”であっても、“nakara”“nagala”“nakala”などと発音される場合がある。このように、似通った音素については、グループ化したパスを持つ隠れマルコフモデルを用いることにより、柔軟性のあるアライメントを実現することができる。

【００５６】[３－３．その他]上記実施形態においては、アライメント対象となる対象音声と入力音声とを時系列で対応づける音声処理装置を、物まね機能を有するカラオケ装置に適用しているが、これに限らず、カラオケ装置であれば例えば採点に用いてもよいし、歌唱を補正するために用いても良い。また、音素単位で時系列を一致させる技術はカラオケ装置に限らず、他の音声認識に関する装置にも適用することが可能である。

【００５７】上記実施形態では、音声信号の代表的な特徴パラメータを特徴ベクトルとして所定数のシンボルにクラスタ化した符号帳と、各音素毎に状態遷移確率および前記各シンボルの観測確率とを記憶する音素辞書について説明しているが、上述した５種類の特徴ベクトルに限らず他のパラメータを用いてもよい。

【００５８】上記実施形態では、対象音声および入力音声をフレーム単位で周波数分析しているが、分析の手法は上述したＳＭＳに限定されるものではないし、時間領域の波形データとして分析しても構わない。あるいは、周波数と波形とを併用した分析をおこなっても構わない。

【００５９】

【発明の効果】以上説明したように、本発明によれば、アライメント対象となる対象音声と入力音声とを時系列で対応づける音声処理を少ない記憶容量でリアルタイムで処理可能となる。

【図面の簡単な説明】

【図１】　本発明の概要を説明する図である。

【図２】　実施形態の構成を示すブロック図である。

【図３】　符号帳を説明する図である。

【図４】　音素を説明する図である。

【図５】　音素辞書を説明する図である。

【図６】　ＳＭＳ分析を説明する図である。

【図７】　対象音声のデータについて説明する図である。

【図８】　実施形態の動作を説明するフローチャートである。

【図９】　入力音声の分析について説明する図である。

【図１０】　隠れマルコフモデルを説明する図である。

【図１１】　アライメントについて具体例を示した図である。

【図１２】　楽曲との同期について説明する図である。

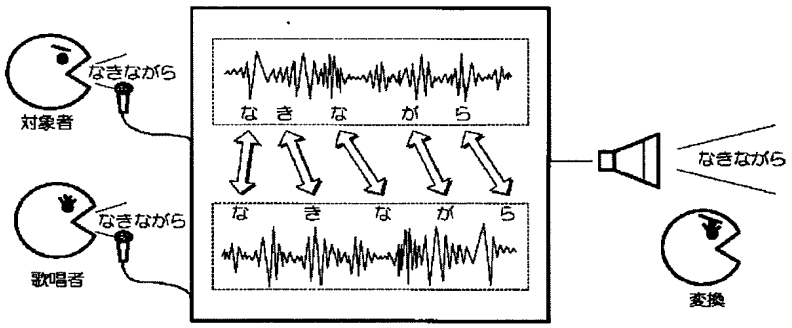
【図１３】　音素の飛び越しを行う場合について説明する図である。

【図１４】　似通った音素が発声される場合について説明する図である。

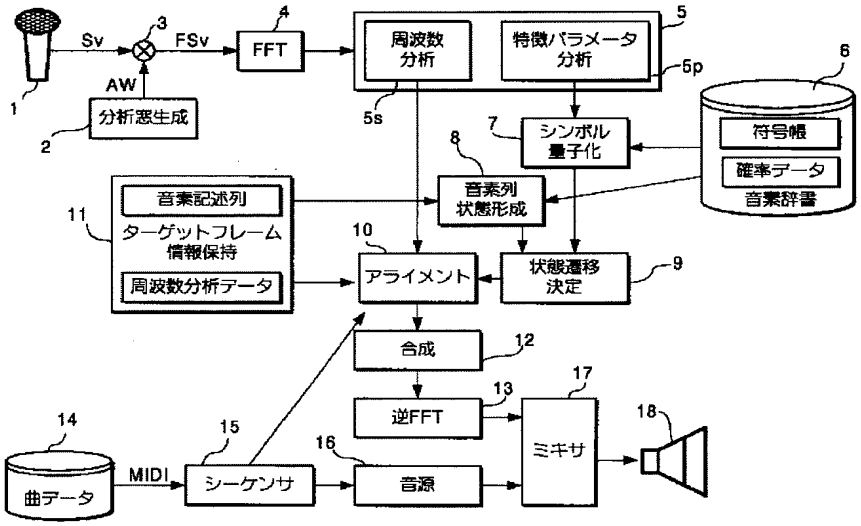
【符号の説明】

１……マイク、
２……分析窓生成部、
３……入力音声信号切出部、
４……高速フーリエ変換部、
５……入力音声分析部、
５ｓ……周波数分析部、
５ｐ……特徴パラメータ分析部、
６……音素辞書記憶部、
７……シンボル量子化部、
８……音素列状態形成部、
９……状態遷移決定部、
１０……アライメント部、
１１……ターゲットフレーム情報保持部、
１２……合成部、
１３……逆高速フーリエ変換部、
１４……曲データ記憶部、
１５……シーケンサ、
１６……音源、
１７……ミキサ、
１８……スピーカ。

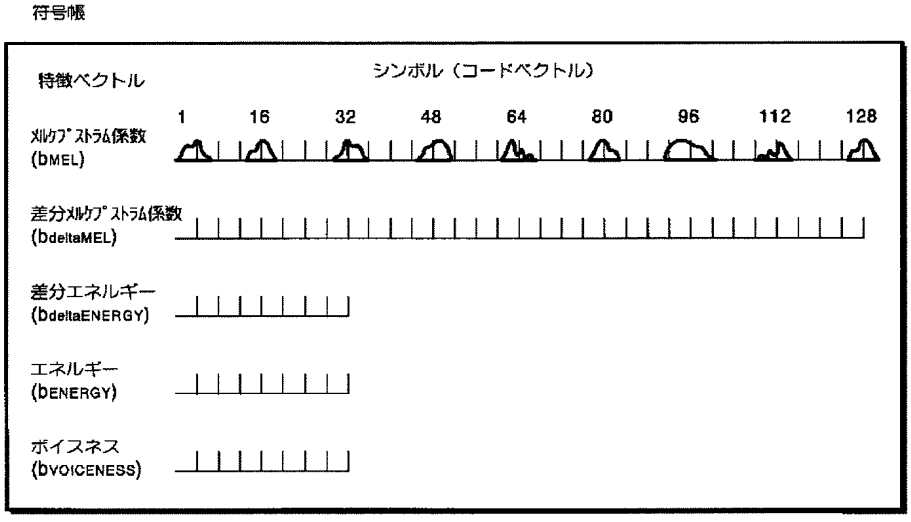
【図１】 Ｆｉｇ．１



【図２】 Ｆｉｇ．２



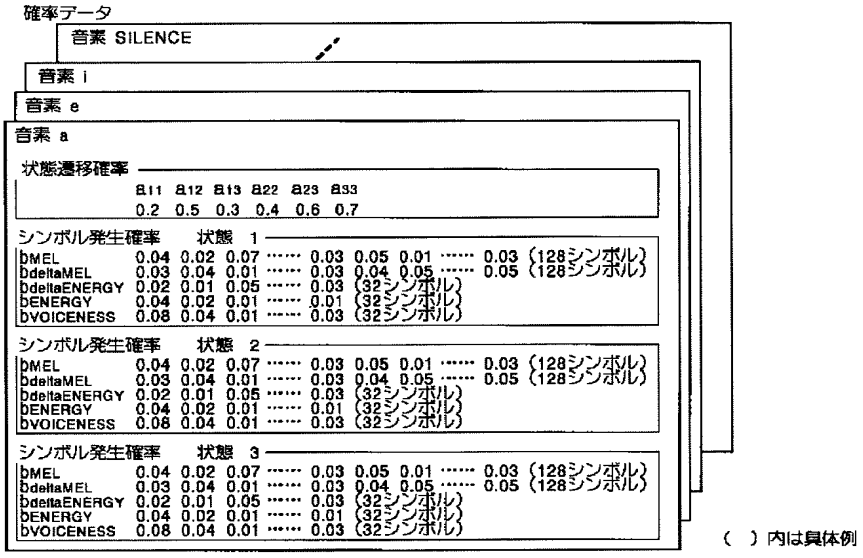
【図３】 Ｆｉｇ．３



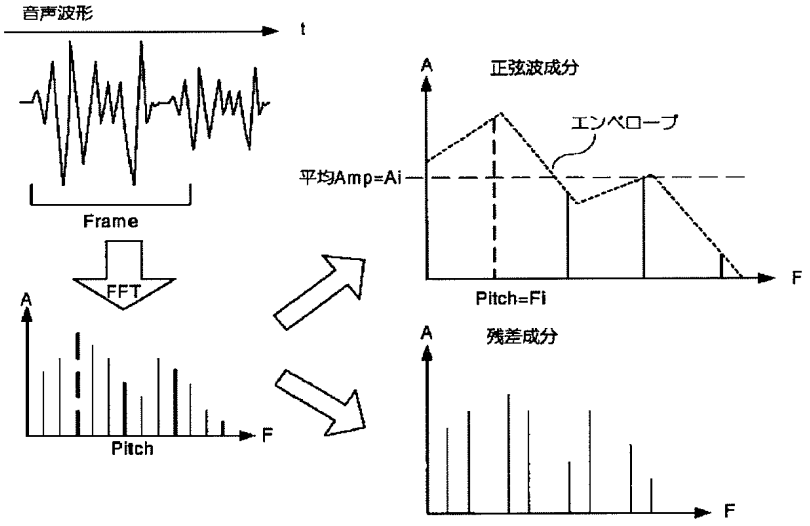
【図４】Ｆｉｇ．４

	音素	状態数
母音	a	3
	e	3
	i	3
	o	3
	u	3
鼻音	m	3
	n	3
	N	3
破裂音	p	2
	b	2
	g	2
	d	2
	t	2
	k	2
摩擦音	s	3
	sh	3
	h	3
	z	3
	ch	3
	ts	3
震え音	r	3
側音	l	3
半母音	w	3
	y	3
二重子音	Q	3
その他	ASPIRATION	2
	SILENCE	1

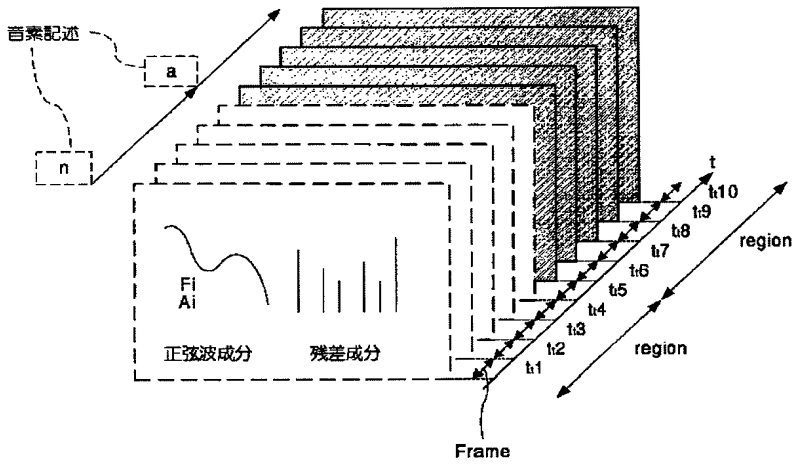
【図５】 Ｆｉｇ．５



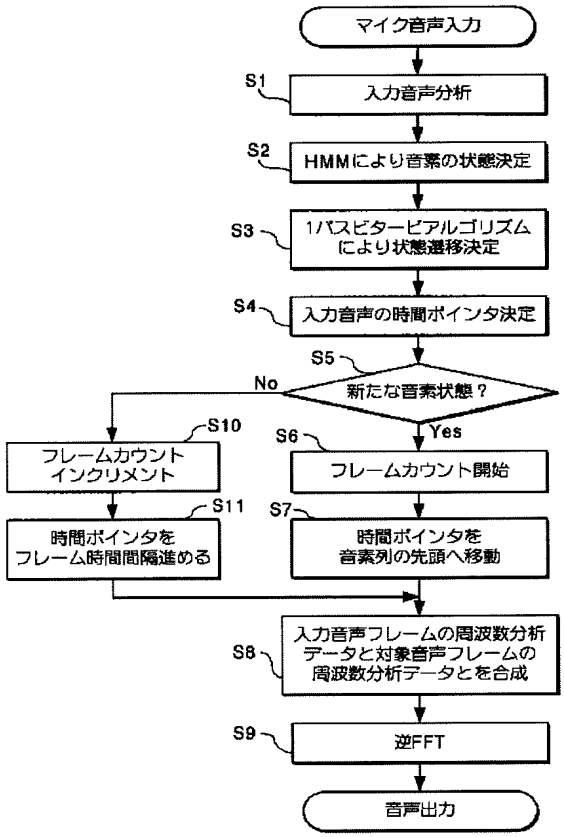
【図６】 Ｆｉｇ．６



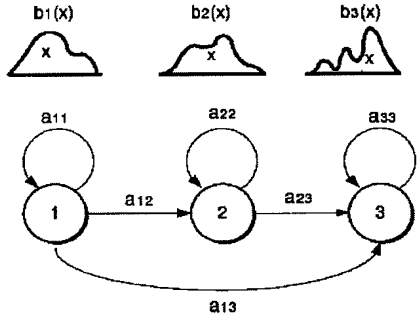
【図７】 Ｆｉｇ．７



【図８】 Ｆｉｇ．８



【図１０】 Ｆｉｇ．１０

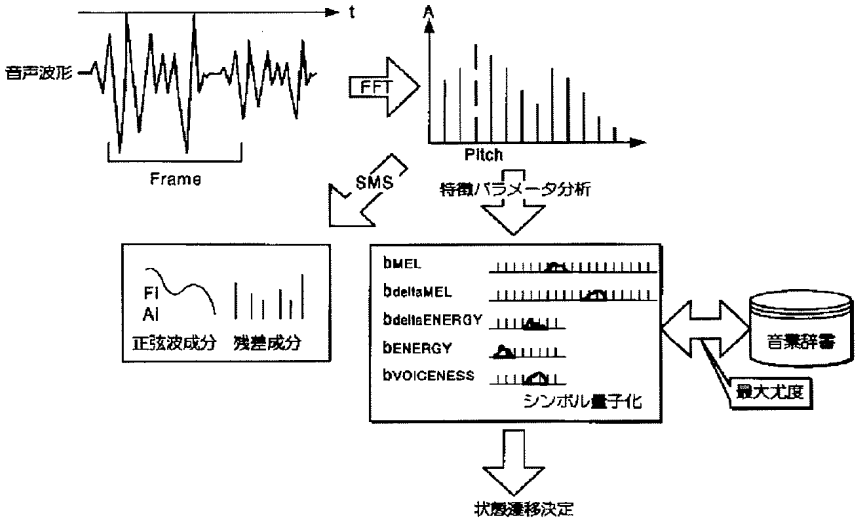


状態遷移確率 a_{ij}

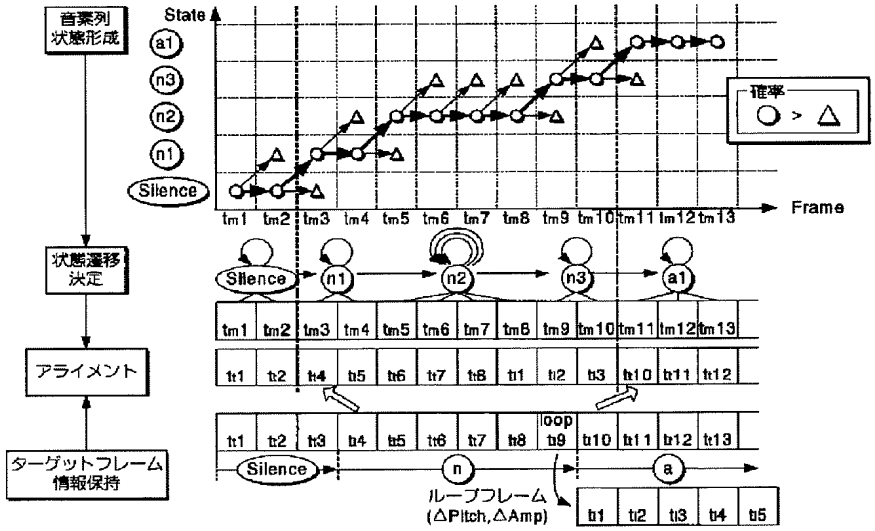
観測シンボル系列 $X=\{x_1,x_2,\cdots,x_T\}$

観測シンボル離散確率 $b_j(x_t)$

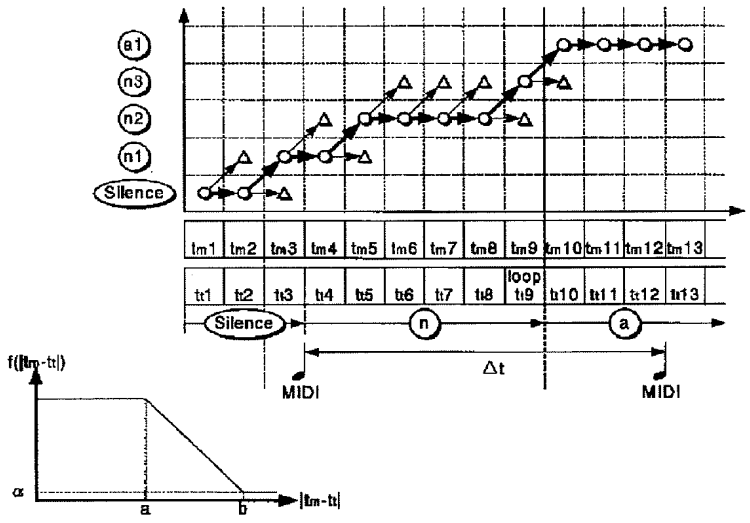
【図９】Ｆｉｇ．９



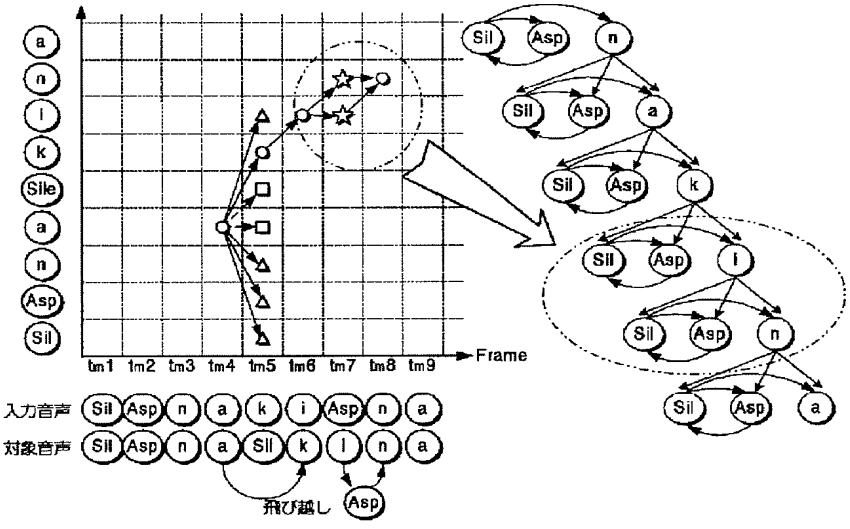
【図１１】Ｆｉｇ．１１



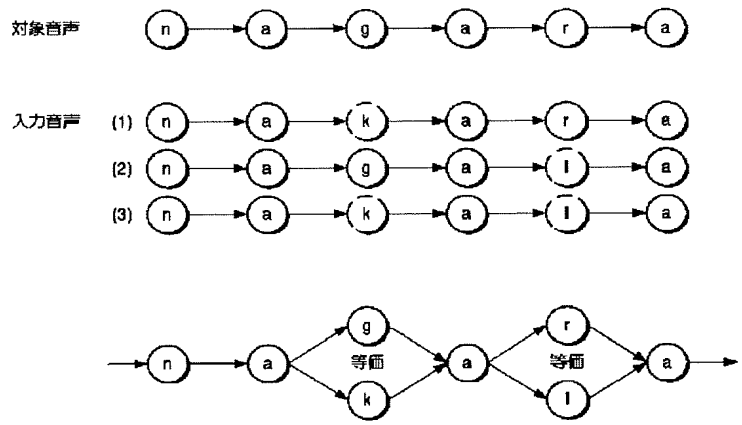
【図１２】Ｆｉｇ．１２



【図１３】Ｆｉｇ．１３



【図１４】 Fig. 14



【手続補正書】
【提出日】平成１１年１１月２６日（１９９９．１１．２６）
【手続補正１】
【補正対象書類名】明細書
【補正対象項目名】００２７
【補正方法】変更
【補正内容】
【００２７】以下、ＬＧＢアルゴリズムを以下に示す。
①初期化
まず、ベクトルの全体の中からセントロイドを見つける。ここでは、初期コードベクトルとする。
②反復
Ｉをトータル反復回数とすると、2^Iのコードベクトルが要求される。そこで、反復回数をｉ＝１，２，・・・・・＊＊，Ｉとすると、反復ｉについて、以下の計算を行う。
１）いくつかの存在するｘというコードベクトルを、ｘ（１＋ｅ）とｘ（１－ｅ）という二つのコードへ分割する。ここで、ｅは、例えば０．００１という小さな数値である。これにより、2^i個の新しいコードベクトルx^i_k（ｋ＝１，２，…，2^i）が得られる。
２）学習セットの中の各々の予測ベクトルｘについて、ｘからコードへx^i_k量子化する。
$k' = argmin_k d(x, x^i_k)$
ここで、$d(x, x^i_k)$は、予測空間での歪み距離を示している。
３）反復計算の間、各々のｋについて、$x^i_k = Q(x)$のように、すべてのベクトルをセントロイドする計算を行う。

【手続補正書】
【提出日】平成１２年２月８日（２０００．２．８）
【手続補正１】
【補正対象書類名】明細書
【補正対象項目名】００４６
【補正方法】変更
【補正内容】
【００４６】［２－２－３－１．１パスビタービアルゴリズム］ビタービアルゴリズムは、観測シンボル系列の各観測シンボルが各ＨＭＭモデルによって出現する全ての確率を算出し、最大確率を与えるパスを後から選択して状態遷移結果とするものである。しかしながら、観測シンボル系列が終結した後に状態遷移結果を求めるので、リアルタイム処理には不向きである。そこで、本実施形態では、以下に説明する１パスビタービアルゴリズムを用いて、その時点まで音素状態を決定する。下記式におけるΨ$_t$（ｊ）は、時刻ｔフレームまでの観測をふまえて算出した、一つのパスを経由して得られる時刻ｔのフレームにおけるベスト確率δ$_t$（ｊ）を最大とする状態を選択する。すなわち、Ψ$_t$（ｊ）に従って音素状態が遷移していく。初期演算としてδ$_1$（ｉ）＝１とし、繰り返し演算として
【数８】

$$\delta_t(j) = \max_{j-1 \le i \le j} \left[\delta_{t-1}(i) a_{ij}\right] \cdot b_{j(MEL)}(O_t) \cdot b_{j(deltaMEL)}(O_t) \cdot b_{j(deltaENERGY)}(O_t) \cdot b_{j(VOICENESS)}(O_t) \cdot b_{j(ENERGY)}(O_t)$$

$$1 \le t \le T, 1 \le j \le t$$

$$\psi_t(j) = \arg\max_{j-1 \le i \le j} \left[\delta_{t-1}(i) a_{ij}\right]$$

$$1 \le t \le T, 1 \le j \le t$$

を実行する。ここで、a_{ij}は状態ｉから状態ｊへの状態遷移確率であり、Ｎは歌唱する曲の音韻数によって決まる状態ｉ、ｊのとりうる最大の状態数である。また、$b_j(O_t)$は特徴ベクトルの時刻ｔにおけるシンボル発生確率である。各観測シンボルは、入力音声から抽出された特徴ベクトルであるから、歌唱者の発声態様によって観測シンボルが異なり、遷移の態様も異なるようになる。図１１に示す例では、上記式によって計算された確率を○あるいは△で示している（○＞△）。例えば、時刻ｔｍ１から時刻ｔｍ３までの観測をふまえ、状態“Ｓｉｌｅｎｃｅ”から状態“ｎ１”へのパスが形成される確率は、状態“Ｓｉｌｅｎｃｅ”から状態“Ｓｉｌｅｎ＊ｃｅ”へのパスが形成される確率よりも高く、時刻ｔｍ３におけるベスト確率となり、図中太矢印で示すように状態遷移を決定する。このような演算を入力音声の各フレームに対応する時刻（ｔｍ１、ｔｍ２、……）毎に行うことによって、図１１に示す例では、時刻ｔｍ３において状態“Ｓｉｌｅｎｃｅ”から状態“ｎ１”に遷移し、時刻ｔｍ５において状態“ｎ１”から状態“ｎ２”に遷移し、時刻ｔｍ９において状態“ｎ２”から状態“ｎ３”に遷移し、時刻ｔｍ１１において状態“ｎ３”から状態“ａ１”に遷移したように決定されている。これにより、入力音声の音素をフレーム単位の各時刻において特定できるようになる。

フロントページの続き

(51)Int.Cl.'	識別記号	FI		テーマコード'（参考）
// Ｇ１０Ｌ	101:04	Ｇ１０Ｌ	3/00	５５１Ｇ
	101:08		5/06	Ｂ
			9/08	３０１Ｂ
			9/14	Ｌ
			9/16	３０１Ｂ

(72)発明者　ペドロ　ケイノ
スペイン　バルセロナ　08002　メルセ12

(72)発明者　アレックス　ロスコス
スペイン　バルセロナ　08002　メルセ12

Ｆターム(参考)　5D015 BB02 CC03 CC06 CC11 GG01 HH11 KK01 KK04
5D045 AB30 DA11
5D108 BF02 BF20